Sharanappa S Yarnal, Ajay N, Shrihari M R / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 3, Issue 4, Jul-Aug 2013, pp.2444-2448

Context Based Dynamic Verbal Management System

Sharanappa S Yarnal[1], Ajay N[2], Shrihari M R[3]
[1, 2, 3] Dept.of Computer Science and Engineering, S.J.C.I.T, Chickaballapur,

Abstract

Verbal ability refers to a person's facility at putting ideas into words, both oral and written. This facility involves possessing not only a strong working vocabulary but also the ability to choose the right words to convey nuances of meaning to a chosen audience. Verbal ability is usually demonstrated as the ability to write and speak well. We describe an approach to construct a learning tool that makes a complex text document to be simple without losing its context and meaning to user or the reader of the document. The word resource that will be using is WordNet which is an open source under Princeton University which is a outcome of 30years of research and dedication. This tool not only make the text to be simple but even this will make the tool to go dynamic by developing its knowledge base with the WordNet and things to go updated into synsets. In this paper we present the methodology for Word Sense Disambiguation based on domain information. Domain is a set of words in which there is a strong semantic relation among the words.

Index Terms–WordNet, Comprehension Index (CI), Word Sense Disambiguation (WSD), Word Knowledge, Domain.

I. Introduction

The Comprehension Index is the index to know the candidates how much understand the words in a different way. The Comprehension Index includes four tests. First, Similarities: Abstract verbal reasoning (e.g., "In what way are an apple and a pear alike?"). Second, Vocabulary: The degree to which one has learned, been able to comprehend and verbally express vocabulary (e.g., "What is a guitar?"). Third, Information: Degree of general information acquired from culture (e.g., "Who is the president of Russia?"). Fourth, Comprehension [Supplemental]: Ability to deal with abstract social conventions, rules and expressions (e.g., "What does *Kill 2 birds with 1 stone* metaphorically mean?"). In the REAP[10] system automatically provides users with individualized authentic texts to read. These texts, usually retrieved from the Web, are chosen to satisfy several criteria. First, they are selected to match the reading level of the student (Collins- Thompson and Callan, 2004). They must also have vocabulary terms known to the student. To meet this goal, it is necessary to construct an accurate model of the student's vocabulary knowledge (Brown and Eskenazi, 2004). Using this model, the system can locate documents that include a given percentage (e.g., 95%) of words that are known to the student. The remaining percentage (e.g. 5%) consists of new words that the student needs to learn. This percentage is controlled so that there is not so much stretch in the document that the student cannot focus their attention on understanding the new words and the meaning of the text. After reading the text, the student's understanding of new words is assessed. The student's responses are used to update the student model, to support retrieval of future documents that take into account the changes in student word knowledge.

Word Sense Disambiguation (WSD) is the process of resolving the meaning of a word unambiguously in a given natural language context. Given a polysemous word in running text, the task of WSD involves examining contextual information to determine the intended sense from a set of predetermined candidates[1]. WSD is task of classification in which the senses are the classes, the context provides the evidence and each occurrence of the word is assigned to one or more of its possible classes based on evidence [2]. The problem is so difficult that it was one of the reasons why the Machine Translation systems were abandoned. However after 1980 large-scale lexical resources and corpora became available and WSD drew attention of researchers. At present WSD is well addressed issue and has occupied important stage in the Natural Language Processing (NLP).

The sense of a word in a text depends on the context in which it is used. The context is determined by the other words in the neighborhood in the sentence. Thus if the word file, hard disk or data appears near the word virus, we can say that it is the program and not the biological virus. This is called as local context or sentential context. One of the first attempts to use dictionary-based approach was by Lesk [3]. He devised an algorithm that chooses the appropriate sense of a polysemous word by calculating the word overlap between the context sentence of the word in question and the word's definition in a Machine Readable Dictionary (MRD).

The Lesk algorithm can be effectively used with the WordNet lexical database. Such attempt is made at IITB [4] and the results are promising. Similar experiments are made by Jonas at Lund University. From the pre-processed documents five words preceding to the word to be disambiguated and five words following it were extracted. These words included nouns, verbs or adjectives. Every sense of

the word to be disambiguated was compared to each sense of surrounding words. Each combination was assigned a score which is based on number of overlapping words. This approach however suffers from the fact that large number of fine senses in WordNet is not distinguishable from each other.

In addition to the importance of these assessments in the REAP system, tests of word knowledge are central to research on reading and language and are of practical importance for student placement and in enabling teachers to track improvements in word knowledge throughout the school year. And also these tools are designed to capture the graded and complex nature of word knowledge, allowing for more fine-grained assessment of word learning.

II. WordNet

WordNet[7][12] is a lexical resource in which English nouns, verbs, adjectives, and adverbs are grouped into synonym sets. A word may appear in a number of these synonym sets, or synsets, each corresponding to a single lexical concept and a single sense of the word (Fellbaum ed., 1998). The word "bat" has ten distinct senses and thus appears in ten synsets in WordNet. Five of these senses correspond to noun senses, and the other five correspond to verb senses. The synset for the verb sense of the word which refers to batting one's eyelashes contains the words "bat" and "flutter", while the synset for the noun sense of the word which refers to the flying mammal contains the words "bat" and "chiropteran". Each sense or synset is accompanied by a definition and, often, example sentences or phrases. A synset can also be linked to other synsets with various relations, including synonym, antonym, hypernym, hyponym, and other syntactic and semantic relations (Fellbaum ed., 1998). For a particular word sense, we programmatically access WordNet to find definitions, example phrases, etc.

In order to retrieve data from WordNet[13], the tool choose the all sense of the word. The system can work with input of varying specificity. The most specific case is when we have all the data: the word itself and a number indicating the sense of the word with respect to WordNet's synsets. When the target words are known beforehand and the word list is short enough, the intended sense can be hand-annotated. More often, however, the input is comprised of just the target word and its part of speech (POS). It is much easier to annotate POS than it is to annotate the sense.

2.2 WordNet Domains

WordNet Domains is an extension of WordNet where synonym set have been annotated with one or more subject domain labels[7]. The domain set used in WORDNET DOMAINS has been extracted from the Dewey Decimal Classification [8] and a mapping between the two taxonomies has been computed in order to ensure completeness.

A domain may include synsets of different syntactic categories: for example MEDICINE groups' together senses from nouns, such as doctor and hospital and from verbs such as operate. A domain may include senses from different WordNet sub-hierarchies.

Domains may group senses of same word into homogeneous clusters. Domains may be used to group together senses of particular word that have the same domain labels. Grouping the words for a particular
Sense reduces the ambiguity while disambiguating it on domain criteria. Table 1 shows the WordNet senses and domains for the word "bank". Table 2 shows the domain distribution over WordNet synsets.

Table 1. WordNet senses and domains for the word "bank".

Sense	Synset & Gloss	Domain
#1	Depository financial institution, bank, banking concern, banking company	ECONOMY
#2	Bank (sloping land ….)	GEOGRAPHY, GEOLOGY
#3	bank (a supply or stock held in reserve…)	ECONOMY
#4	bank, bank building (a building…)	ARCHITECTURE, ECONOMY
#5	bank (an arrangement of similar objects….)	FACTOTUM
#6	saving bank, coin bank, money box. bank (a container)	ECONOMY
#7	bank (a long ridge or pie…)	GEOGRAPHY, GEOLOGY
#8	Bank (the funds held by gambling house….)	ECONOMY, PLAY
#9	Bank, cant, camber (a slope in the turn of road…)	ARCHITECTURE
#10	Bank (a flight maneuver…)	TRANSPORT

Table 2. shows the Domain distribution over WordNet synsets

Domain	#Syn	Domain	#Syn	Domain	#Syn
Factotum	36820	Biology	21281	Earth	4637
Psychology	3405	Architecture	3394	Medicine	3271
Economy	3039	Alimentation	2998	Administration	2975
Chemistry	2472	Transport	2443	Art	2365
Physics	2225	Sport	2105	Religion	2055
Linguistics	1771	Military	1491	Law	1340
History	1264	Industry	1103	Politics	1033
Play	1009	Anthropology	963	Fashion	937
Mathematics	861	Literature	822	Engineering	746
Sociology	679	Commerce	637	Pedagogy	612
Publishing	532	Tourism	511	Computer_Science	509
Telecommunications	493	Astronomy	477	Philosophy	381
Agriculture	334	Sexuality	272	Body_Care	185
Artisanship	149	Archaeology	141	Veterinary	92
Astrology	90				

III. Algorithm

To disambiguate a word, three types of bags are used: The algorithm tags the text with part of speech tags using pos tagger. Bag b1 contains the content words. The domains for each word from b1 relating to pos tag sense are selected and inserted into bag b2. The domains corresponding to pos tag of target word are Inserted into bag b3. For each domain in b3, the domains of the other words are compared (domain factotum can match with any domain). The domain of target word which maximises the match with the domains of other words becomes the domain of the text. The sense belonging to this domain is the correct sense of the target word.

Let us assume that (*w1, w2, w3,..wn*) is the bag *b1* containing pos tagged content words. And b2 is the bag containing (*d1, d2, d3,…dn*), sets of all domains corresponding to the content words w.r.t their pos tags. Each set di contains all possible domains corresponding to the pos tag sense. The bag b3 contains the domains of target word.

1. Input the sentence
2. Perform POS Tagging.
3. From the POS tagged text separate the content words and insert into bag b1.
4. For each content word, insert set of domains corresponding to it's pos tag into bag b2.
5. For the target word wt, insert domains corresponding to it's pos tag into bag b3
6. Compare each domain in b3 with set of domains of remaining content words.
7. The domain in b3 which maximises with domains of other content words is the domain of the text.
8. The sense belonging to domain obtained from step 7 is the correct sense.

3.1 Example

Let us consider the sentence

The *virus* infected all files on the disk.

This sentence is passed to the POS Tagger. The output is the tagged text as follows
The/DT virus/NN infected/VBD all/DT files/NNS on/IN disk/NN ./.

Out of the above the system selects only the content words viz. virus, infected, files and disk. Suppose we want to find the correct sense of the word *virus* in the above sentence. The sense numbers and domains of target word and that of content words are shown in Figure 1.

Virus (noun sense)
Infected (verb sense)
Files (noun sense)
disk (noun sense)

b1

Virus (noun sense) Target word	Infected (verb sense)	Files (noun sense)	disk (noun sense)
01254816-factotum	00087224-medicine	06106818-telecommunications	03364489-computer science
13209397-factotum	00086241-medicine	07917489-factotum	
06179311-Computer_science	02503346-factotum	03215630-administration furniture	
	00585683-phychological features	03215329-building industry	

b2

01254816-factotum
13209397-factotum
06179311-Computer_science

b3

Figure 1. Contents of bag b1, b2 and b3.

In the given example
b1= {virus, infected, files, disk}
b2= {{medicine, medicine, factotum, psychological_features},{telecommunication, factotum, administration, building industry}, {computer science}}

wt= virus

b3= {factotum, factotum, computer science}

Since the domain computer-science from bag b3 has maximum score (factotum, factotum, computer science), we fix computer-science as the domain of the
text. The noun sense belonging to computer science domain is the selected as the correct sense.

IV. SMOG

The **SMOG grade** is a measure of readability that estimates the years of education needed to understand a piece of writing. SMOG is the acronym derived from Simple Measure of Gobbledygook. It is widely used, particularly for checking health messages. The SMOG grade yields a 0.985 correlation with a standard error of 1.5159 grades with the grades of readers who had 100% comprehension of test materials.

The formula for calculating the SMOG grade was developed by G. Harry McLaughlin as a more accurate and more easily calculated substitute for the Gunning fog index and published in 1969. To make calculating a text's readability as simple as possible an approximate formula was also given — count the words of three or more syllables in three 10-sentence samples, estimate the count's square root (from the nearest perfect square), and add 3.

Table 3: SMOG Conversion Table

SMOG Conversion Table	
Total Polysyllabic Word Count	**Approximate Grade Level (+1.5 Grades)**
1 – 6	5
7 – 12	6
13 – 20	7
21 – 30	8
31 – 42	9
43 – 56	10
57 – 72	11
73 – 90	12
91 – 110	13
111 – 132	14
133 – 156	15
157 – 182	16
183 – 210	17
211 – 240	18

1.1 Formulae

To calculate SMOG

1. Count a number of sentences (at least 30)
2. In those sentences, count the polysyllables (words of 3 or more syllables).
3. Calculate using

SMOG grade = 3 + Square Root of Polysyllable Count

This version (sometimes called the SMOG Index) is more easily used for mental math:

1. Count the number of polysyllabic words in three samples of ten sentences each.
2. Take the square root of the nearest perfect square
3. Add 3

SMOG conversion tables as shown in table3.

V. Analysis

The meaningful sentences are composed of meaningful words; any system that hopes to process natural languages as people do must have information about words and their meanings. This information is traditionally provided through dictionaries, and machine-readable dictionaries are now widely available. But dictionary entries evolved for the convenience of human readers, not for machines. WordNet provides a more effective combination of traditional lexicographic information and modern computing. WordNet is an online lexical database designed for use under program control. English nouns, verbs, adjectives, and adverbs are organized into sets of synonyms, each representing a lexicalized concept. Semantic relations link the synonym sets.

Microsoft Word is the de facto word processor. Whether we work at home, at school or in business, the chances are that we will use MS Word if you need to create your own, or read someone else's document. Word can seem a little frightening at first - especially if you are coming to Word 2007 from previous versions. Word thesaurus can take advantage of to improve our documents. Thesaurus can be used to find synonyms (different words with the same meaning) and antonyms (words with the opposite meaning).Microsoft Word displays the synonym list in the form of most frequently used to least frequently use. For example, we can take a word

as 'develop', the synonyms are expand, build up, enlarge, extend, increase, widen, and grow. But when compared with the WordNet browser, the synonyms are displayed depends on the senses of word.

VI. Conclusion

The paper is described to manipulate the text document to produce user readable text document using English grammar rules and replacing by the available nouns for the polysyllables, without losing the meaning and the context and also to find the comprehension index of the reader and text document using SMOG readability formulae.

It has been observed that most of the synsets(36820) in WordNet belong to the domain factotum and do not contribute for any domain information. Further a word may have multiple senses for a particular domain. For example the word bank has sense#1, sense#6 and sense#8 belonging to the domain economy as shown in Table 1. Thus, all methods based on simple frequency counting often turn out to be inadequate. The drawback of the algorithm is that it can disambiguate a word provided it has only one sense per domain.

References

[1] Chen, J. and J. Chang 1998. *Topical Clustering of MRD Senses Based on Information Retrieval techniques.* Computational Linguistics. MIT Press, Cambridge, MA. Vol.24(1), pp. 61-95.

[2] E. Agirre and G. raigu. 1996. *Word Sense Disambiguation using Conceptual Density.* In Proceeding of COLLING, pages 16-22.

[3] M. Lesk, *Vocabulary problems in retrieval systems*, in Proc. 4th Annual Conference of the University of Waterloo Centre for the New OED. 1988.

[4] Ganesh Ramakrishnan, B. Prithviraj, Pushpak Bhattacharyya. *A Gloss Centered Algorithm for Word Sense Disambiguation.* Proceedings of the ACL SENSEVAL 2004, Barcelona, Spain. P. 217-221.

[5] Magnini, B., Strapparava, C., Pezzulo, G., Gliozzo, A., 2002. *The role of domain information in word sense disambiguation.* Natural Language Engineering 8 (4), 359–373.

[6] Gregory Aist. 2001. Towards automatic glossarization: automatically constructing and administering vocabulary assistance factoids and multiple-choice assessment, International Journal of AI in Ed., 2001.

[7] Miller G. A., Beckwidth, R., Fellbaum, C., Gross, D. and Miller, K. J., "Introduction to WordNet: An On-line Lexical Database", International Journal of Lexicography, Vol 3, No.4 (Winter 1990), pp. 235-244.

[8] Jonathan C. Brown, Gwen A. Frishkoff, Maxine Eskenazi, "Automatic Question Generation for Vocabulary Assessment", Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 819–826, Vancouver, October 2005.

[9] S. G. Kolte, S. G. Bhirud, "Word Sense Disambiguation using WordNet Domains", First International Conference on Emerging Trends in Engineering and Technology, IEEE, 2008.

[10] Myunggwon Hwang, Byungsu Youn, Ilyong Chung, Pankoo Kim, "Semantic Measurement of Related degree between Unknown Word and Related Word for Automatic Extension of Lexical Dictionary", Fifth International Conference on Fuzzy Systems and Knowledge Discovery, IEEE, 2008.

[11] Kevyn Collins-Thompson and Jamie Callan. 2004. A language modeling approach to predicting reading difficulty. In Proceedings of the HLT/NAACL 2004 Conference. Boston, 2004.

[12] http://wordnet.princeton.edu /edu.mit.jwi_2.1.5_manual

[13] Stephanie Chua, Narayanan Kulathuramaiyer, "Semantic Feature Selection Using WordNet", Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04), 2004.

[14] Raghuvar Nadig J. Ramanand1 Pushpak Bhattacharyya, "Automatic Evaluation of Wordnet Synonyms and Hypernyms", Proceedings of the 6th International Conference on Natural Language Processing, 2008.

[15] *Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, Maxine Eskenazi, "*Classroom Success of an Intelligent Tutoring System for Lexical Practice and Reading Comprehension*"* Language Technologies Institute Carnegie Mellon University, Pittsburgh, Pennsylvania, United States of America,2004

Mr. Sharanappa S Yarnal is currently pursuing final year Bachelor of Engineering in Computer Science and Engineering Department at S J C Institute of Technology, Chickballapur, Affiliated to Visvesvaraya Technological University, Belgaum, Karnataka, India.

Mr. Ajay N is an Assistant professor in the Department of Computer Science and Engineering at S J C Institute of Technology ,Chickballapur affiliated to Visvesvaraya Technological University, Belgaum, Karnataka, India.

Mr. Shrihari M R is an Assistant professor in the Department of Computer Science and Engineering at S J C Institute of Technology, Chickballapur , affiliated to Visvesvaraya Technological University, Belgaum, Karnataka, India.